PB-237 806-T

INTERPRETATION OF LOW-RESOLUTION MASS SPECTRA BY CLASSIFICATION WITH NEIGHBOR SPECTRA FROM A LIBRARY. DETERMINATION OF SUBSTANCE CLASSES AND CHEMICAL STRUCTURES

K. Varmuza

Environmental Protection Agency
Research Triangle Park, North Carolina

1974

DISTRIBUTED BY:

NTIS
National Technical Information Service
U. S. DEPARTMENT OF COMMERCE

312035

NTIS

PB 237 806-T

EPA-TR 491-74

K. Varmuza

INTERPRETATION OF LOW-RESOLUTION MASS SPECTRA BY CLASSIFICATION WITH NEIGHBOR SPECTRA FROM A LIBRARY

Determination of Substance Classes and Chemical Structures

[Interpretation niedrig aufgeloester Massenspektren durch Klassifikation mit Hilfe der Nachbarspektren aus einer Bibliothek. Ermittling von Substanzklassen und chemischen Strukturen.]

Z. Anal. Chem., Vol. 268, No. 5, 1974, pp. 352-356.

Reproduced by
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. Department of Commerce
Springfield VA 22151

Translated for the
ENVIRONMENTAL PROTECTION AGENCY
by
Translation Consultants, Ltd., Arlington, Virginia

17

K. Varmuza

INTERPRETATION OF LOW-RESOLUTION MASS SPECTRA BY CLASSIFICATION WITH NEIGHBOR SPECTRA FROM A LIBRARY. Determination of Substance Classes and Chemical Structures

Summary

Interpretation of Low Resolution Mass Spectra by Classification with Neighbor-Spectra from a Library. Determination of Molecular Structures. The Euclidean distance between mass spectra in an n-dimensional hyperspace is a good measure of chemical-structural similarity of low molecular organic substances. The KNN-method, well-known in pattern recognition, was used to find neighbor-spectra from a library which are close to the mass spectrum of an unknown substance. These neighbors are very suitable to determine some classes of molecular structures of the unknown substance, which need not be a member of the library. On an average the first neighbor has the highest classifying capability.

Introduction

The automatic identification of substances with the aid of low-resolution mass spectra is, at present, done primarily through comparison of the sample spectrum with a spectra data bank. As far as the degree of conformity of two mass spectra is concerned, the most divergent similarity criteria were suggested; in most instances, they took into consideration only few peaks of the spectra and endeavored to establish as rapid a search procedure as possible [5,6,12]. If the unknown spectrum is not contained in the data bank, it is assumed that the search procedure will turn up spectra of "similar" substances. Nothing is as yet known concerning the extent

of said similarity. Thus, the question arises, for instance, as to what chemical structures in an unknown substance can, with any degree of accuracy, be ascertained in this manner.

This report deals for the first time with the results of a search procedure for the interpretation of mass spectra, which has long been known in the field of automatic pattern recognition and which has recently been suggested for the evaluation of nuclear resonance spectra [7]. In accordance with this so-called KNN-method (k-nearest neighbor classification rule), one uses as similarity criterion the distance between two spectra in the polydimensional spectra space. Those k data bank spectra, which lie nearest to the unknown spectrum and which hereafter will be referred to as neighbors, are utilized for the classification of the substance category and for the chemical structure elements. Although this search procedure is, on principle, also suitable for mass spectra with a reduced number of peaks, the spectra used here for experimental purposes were complete. This search procedure requires more time than the rapid methods recommended in the past. In many cases, however, the time factor is going to be less significant than the possibility of obtaining a great deal of information from the low-resolution mass spectrum of an unknown substance.

KNN-method

The mass spectra of a spectra data bank may be represented in the form of X_i dots in an n-dimensional space, whereby n indicates the total of mass numbers which may have peaks in the spectra. The coordinates x_{i_1}, x_{i_2}, ..., x_{im}, ... x_{in} of an X_i dot do, for instance, correspond to the peak heights in the mass spectrum i. The distance between points X_i and X_j is indicated through the Euclidian distance d_{ij}, which, analogous to the clearly perceptible two-dimensional or three-dimensional case, is defined as

$$d_{ij} = \sqrt{\sum_{m=1}^{n} (x_{im} - x_{jm})^2}$$

whereby d_{ij} is measured in the same units as the peak heights in the spectra. In order to classify a sample spectrum X_o, in accordance with the KNN-method [1,2,9,10], one selects from among the data bank spectra those k spectra (neighbors) which have the shortest distance to X_o, i.e. have the greatest degree of similarity. For this purpose, it is necessary to compute the distances between X_o and all data bank spectra. The k neighbors, for which the substance categories and chemical structures are known, classify the association between the mass spectrum X_o and the individual categories in accordance with

a synchronization method to be specified. In several instances of applying the KNN-method, it became apparent that the X_o could be best classified by means of the nearest (first) neighbor (1NN-method) [1,3,7]. In the case of more extensive spectra data banks, it would be preferable to have the majority of $k > 1$ neighbors determine the category association. One also ought to take into weighting an assessment of the individual neighbors and a consideration of the distance parameters to X_o. The synchronization result of the k neighbors might possibly also provide an indication of the classification reliability. The KNN-method may also be looked upon as an approximate determination of the probable density of a category at the X_o spot. It has frequently been suggested as a standard method for examining other classification methods [1,7,8].

The interpretation of a mass spectrum requires that one answer several questions relating to the association between the sample, on one hand, and substance categories and chemical structures, on the other hand. Depending upon the synchronization result, the answer is either "yes" (i.e., the sample is classified as belonging to the category in question), "no", or "rejection" of the question. It is noteworthy that if the KNN-method is applied, all category classifications in question referring to a certain X_o spectrum are determined by the same neighbors.

Data and programs

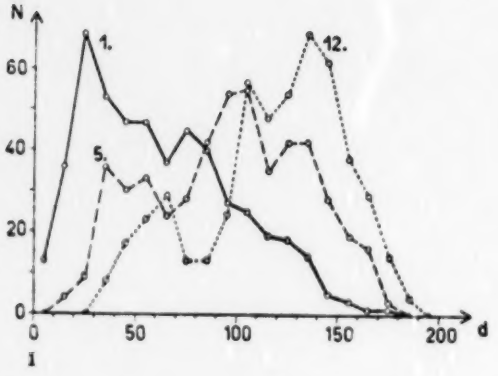
The same collection of the mass spectra of 500 different substances was used, which had earlier [11] been utilized for the computation of adaptive, linear classifiers. The substances have molecular weights ranging between 16 and 150 and chemical formulas of the type $C_{1-11}H_{1-22}O_{0-2}N_{0-2}$. For each substance the association with 31 different categories was determined (substance categories, existence of certain structural elements, oxygen or nitrogen, respectively, in the molecule).

In the following research, each of the 500 mass spectra was perceived as spectrum X_o of an unknown sample and was in each case compared with the 400 remaining data bank spectra; subsequently, the twelve nearest neighbors were selected for each individual case. The arithmetical work was carried out with FORTRAN IV programming on an IBM 7040 in the Computer Center of the Technical University at Vienna. The spectra data bank was stored on tape. The programs permit the simultaneous comparison of 50 unknown mass spectra with the data bank; the computing time for this procedure amounted to about 15 minutes (the programs, though, are in no way maximized with respect to operating time).

Results

Euclidian distance as similarity criterion

Since in the KNN-method here employed the Euclidian distance d_{ij} between two mass spectra X_i and X_j in the spectra space is used as similarity criterion for substances i and j, it must first be determined whether the desired connection exists between d_{ij} and the chemical-structural similarity of i and j. Figure 1 depicts the frequency distributions of the distances to the first, fifth, and twelfth neighbor in the utilized spectra data bank. As anticipated, the maximum of the distribution curve lies for the first neighbor at a considerably lesser distance than for the twelfth neighbor. In the case of some spectra, the distance to the first neighbor is, however, greater than the distances of other spectra to their twelfth neighbor. Obviously, certain portions of the n-dimensional spectra space are covered rather densely with spectra dots, whereas more remotely situated spectra are separated from all other members of the data bank by considerable distances. In view of the relatively small spectra data bank, no further research of the cluster formation was undertaken.



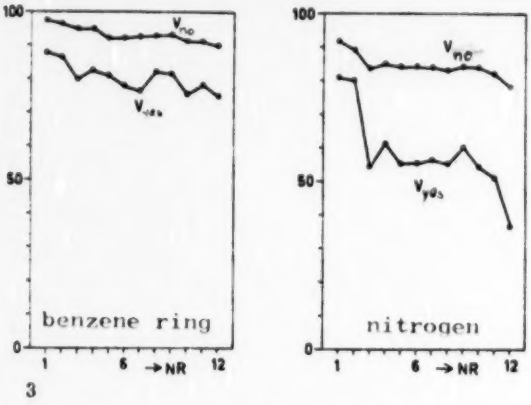


Fig. 3. Reliabilities V_{yes} and V_{no} of the yes/no-answers in the case of classification of a benzene ring or of the nitrogen existence in the molecule, respectively, through a single neighbor of the NR numerical value

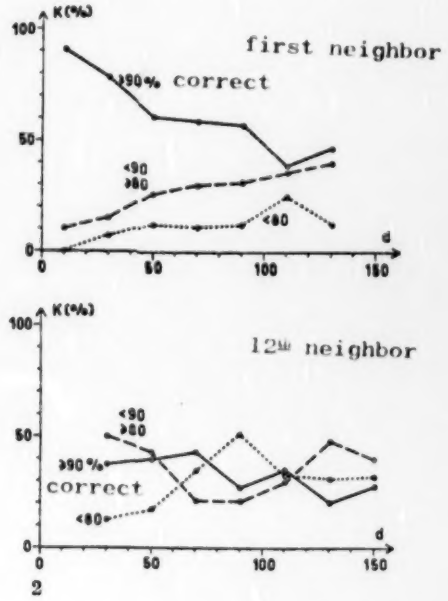


Fig. 1. Frequency distributions of the distances to each first, fifth, and twelfth neighbor. d is the Euclidian distance, subdivided into 20 intervals of width 10; N is the total of mass spectra, for which the first, fifth, and twelfth neighbor falls into the corresponding distance interval

Fig. 2. Classification capability of the first and twelfth neighbor in relation to the Euclidian distance d between the neighbor and the unknown mass spectrum. K is the percentage of those spectra, for which the first (twelfth) neighbor falls into the corresponding distance interval, and for which said neighbor correctly classifies the 31 categories on the specified scale

Subsequently, the classification capability of the individual neighbors is examined in relation to the distance. The question is formulated as follows: Is the first neighbor much better suited for purposes of classification than other neighbors, and is the classification capability dependent upon the distance of the neighbor? For this purpose, 500 spectra were divided into ten groups, so that the distances between the spectra and their k-neighbors would in each case fall into one of the ten distance intervals 0-20, 20-40, ..., 180-200. In each group the percentage of those spectra was determined for which the allocation to the 31 categories, on the basis of the k^{th} neighbor, was correct $> 90\%$; $> 80\%$, but $< 90\%$; and $< 80\%$. Figure 2 shows these distributions for the first and twelfth neighbor (the gradient of the curve follows identical courses from the first to the twelfth neighbor). It appears from the diagrams that the classification capability of the first neighbor is considerably greater than that of the twelfth neighbor. Moreover, the dependency of the classification capability upon the distance is clearly pronounced in the case of the first neighbor; the farther away the first neighbor is situated, the fewer categories of the sample spectrum are, as a rule, correctly allocated. If, for instance, one utilizes in each individual case only the first

neighbor for purposes of classification and rejects all those cases in which the distance to the first neighbor is in excess of 100 units, one obtains the following average values:

Percentage of correctly determined categories (of a total of 31)	Percentage of the mass spectra (of a total of 500)
>90%	56.4%
>80 to < 90%	19.0%
<80%	7.6%
Rejections	17.0%

In this connection it ought to be noted that each spectrum was classified through a spectrum of another substance. In actuality, a mass apectrum of the examined substance is in many instances already going to be contained in the data bank and the results are likely to be considerably more favorable.

Classification according to individual substance categories and chemical structure elements

Up to now the classification capability was summarily examined for all 31 selected categories. However, it is also significant to study the method with respect to its applicability to each individual category. As previously shown [11], one must see to it that a classification procedure correctly allocates a large percentage of those spectra of substances that belong to a category (Category YES), as well as those that do not belong to the category (Category NO). P_{YES} and P_{NO} are the percentages of the

accurately classified spectra of Category YES or Category NO, respectively; these values must be individually determined for each of the 31 categories. The user of a classification procedure is interested primarily in the reliabilities V_{yes} and V_{no} of the yes-no responses.* V_{yes}, for instance, indicates what percentage of the yes-answers, which were given in response to a question concerning a certain category, are on the average accurate.

Figure 3 depicts typical curves for V_{yes} and V_{no}, which characterize the cluster formation of two categories. The registered measuring values were obtained through exclusive use in each case of the first to twelfth neighbors for the classification of a benzene ring, as well as for the classification of nitrogen in the molecule. In both respects the first neighbor is most suitable for classification purposes. The mass spectra of substances with benzene ring constitute a rather compact mass, since the other neighbors as well still supply good results. In the case of nitrogen classification, however, a strong decline in the reliability of the yes-answer sets in after the second neighbor.

---

* In the case of the classification according to "alcohol," all alcohols are contained in the YES category, all other substances in the NO category. In this context, the categories are referred to in capital letters (YES, NO), the responses of the classifier in lower case (yes, no).

Table 1 contains the results which ensue, if in each case only the first neighbor (1NN-method) is used for the classification of the 31 substance categories and chemical structures. The good classification capability for the NO categories and the high degree of reliability for the no-answers are not surprising, since the share of NO categories is very considerable in practically all categories. In the following nine classifications, V_{yes} as well as V_{no} amounted to at least 70%: alkane, alkyl benzol, the CO-group of an aliphatic compound, substances with (substituted) NH_2-groups, aliphatic substances with (substituted) NH_2-groups, substances with benzene rings, phenols, substances with oxygen (or nitrogen, respectively) in the molecule. Since these substance categories, too, are in most cases represented only sparsely in the spectra data bank, it must be assumed that these mass spectra form well-defined clusters which are only slightly interspersed with spectra of other substances. Five of the nine enumerated categories supplied good results, even when adaptive, linear, binary classifiers were employed [11].

Discussion

The Euclidian distance between the mass spectra in the n-dimensional spectra space represents in the case of some substance categories a good criterion for the

chemical similarity of the substances. In these cases, a search procedure based on the KNN-method supplies, even for mass spectra of substances not contained in the data bank, valuable indications relating to the structure of the unknown substance. The distinguishing characteristics of the KNN-method are its specific simplicity and its insignificant programming expenditures; it does not require any training of decision vectors and is generally also suitable for those cases in which a linear classifier, due to a complicated cluster structure, fails to function. The employed spectra data bank can be expanded at any time without requiring any modification of the KNN search program. The considerable time consumed in classifying unknown spectra represents a certain drawback, since it is always necessary to gather the entire data bank from an external memory device and to compute the distances to all data bank spectra. A reduction of the spectra data bank in accordance with the CNN-method (condensed nearest neighbor rule [3,4]) to the required spectra ought to be tested. A decrease in the peak intensities of the mass spectra to only a few phases and a reduction in the number of peaks would also considerably shorten the computation time and thus ought to be taken into consideration. In this report the classification was in each case based only on the first neighbor.

Several tests with classifications through a majority of k neighbors showed definite improvements with respect to the reliability of the yes/no-answers; this is going to be taken up in a later report.

Table 1. Classification according to substance categories and chemical structures on the basis of the 1NN-method. A is the share of the YES category of a certain classification in relation to the entire data bank. V_{yes} and V_{no} are the reliabilities of the yes/no-answers. P_{YES} and P_{NO} are the shares of correctly classified mass spectra of the YES AND NO categories. The precise significance was previously [11] described.

No.	Class	A(%)	V_{yes}(%)	V_{no}(%)	P_{YES}(%)	P_{NO}(%)
1	Alkane	8.0	75.6	98.0	77.5	97.8
2	Cyclic KWST	3.4	57.1	98.2	47.1	98.8
3	Alkenes	5.4	45.7	98.7	77.8	94.7
4	Dienes	1.6	33.3	98.8	25.0	99.2
5	Alkyl benzol	6.0	77.4	98.7	80.0	98.5
6	Aliphatic alcohol	11.8	58.5	93.7	52.5	95.0
7	Ether	10.4	45.8	93.4	42.3	94.2
8	Ketone	8.2	37.0	94.7	41.5	93.7
9	Aliphatic ketone	6.4	40.0	96.1	43.8	95.5
10	Aromatic ketone	1.0	0.0	99.0	0.0	99.0
11	Aldehyde	5.6	43.5	96.2	35.7	97.3
12	Carbonic acid	3.0	40.0	97.8	26.7	98.8
13	Ester	8.4	40.5	94.2	35.7	95.2
14	CO-group	28.0	66.9	84.8	59.3	88.6
15	Aliphatic CO	22.2	70.1	89.3	61.3	92.5
16	Aromatic CO	4.6	59.1	97.9	56.5	98.1
17	NR_2-group	14.2	78.5	95.4	71.8	96.7
18	Amide	2.8	62.5	98.2	35.7	99.4
19	Aliphatic NR_2	12.0	85.5	97.1	78.3	98.2
20	Aromatic NR_2	2.0	14.3	98.2	10.0	98.8
21	Benzene ring	22.6	87.9	97.1	90.3	96.4
22	Phenol	4.2	70.8	99.2	81.0	98.5
23	Pyridine ring	7.0	52.6	96.8	57.1	96.1
24	Methyl group	77.6	85.6	61.9	91.8	46.4
25	Ethyl group	40.2	61.2	75.5	65.2	72.2

Table 1. - Continued

No.	Class	A(%)	V_{yes}(%)	V_{no}(%)	P_{YES}(%)	P_{NO}(%)
26	n-propyl group	22.2	42.7	84.1	45.1	82.8
27	Vinyl group	9.8	41.7	94.6	51.0	92.2
28	C=C double compound	18.2	59.4	92.2	65.9	90.0
29	Branching at the C	24.4	60.5	88.1	63.9	86.5
30	Oxygen	53.4	86.5	78.1	79.0	85.8
31	Nitrogen	24.6	80.5	91.7	74.0	94.2

Acknowledgment

I express my gratitude to Professor Dr. A. Maschka for his kind support in preparing this report. Further thanks are due to Dr. P. Krenmayr for valuable discussions and to Mr. H. Urban for his effective collaboration.

Bibliography

1. Cover, T.M., P.E. Hart, IEEE Trans IT-13, p. 21, 1967.
2. Fix, E., J.L. Hodges, Discriminatory analysis, non-parametric discrimination: Consistency properties, Project 21-49-004, Report 4, Univ. of California, 1951, as quoted by Nilsson [10], p. 126.
3. Gates, G.W., IEEE Trans. IT-18, p. 431, 1972.
4. Hart, P.E., IEEE Trans. IT 14, p. 515, 1968.
5. Hertz, H.S., R.A. Hites, K. Biemann, Anal. Chem., Vol. 43, p. 681, 1971.
6. Knock, B.A., I.C. Smith, D.E. Wright, R.G. Ridley, W. Kelly, Anal. Chem., Vol. 42, p. 1516, 1970.

7. Kowalski, B.R., C.F. Bender, Anal. Chem., Vol. 44, p. 1405, 1972.

8. Kowalski, B.R., C.F. Bender, J. Am. Chem. Soc., Vol. 94, p. 5632, 1972.

9. Meisel, W.S., "Computer-oriented approaches to pattern recognition," New York-London: Acad. Press, 1972.

10. Nilsson, N.J., "Learning machines," New York: McGraw-Hill, 1965.

11. Varmuza, K., P. Krenmayr, This magazine, Vol. 266, p. 274, 1973.

12. Wangen, L.E., W.S. Woodward, T.L. Isenhour, Anal. Chem., Vol. 43, p. 1605, 1971.